Traineeship data analysis methodology

All programming is done in R using the following packages.

Tidyverse – for data cleaning

data.table – For fread to read large data tables

countrycode – To exctract ISO country does

hablar

corrr – For correlation using pipes

janitor – To clean column names to snake\_case

WDI – For world bank data

sf – For loading map data

tmap – For creating maps

# Gathering of data

All data was collected from the FAO database on food production and trade, WDI, an R-package to access world bank data.

# Food production

Datasource: FAO

<http://www.fao.org/faostat/en/#data/QC>

All countries, elements and items are chosen for the year 2016 and data downloaded as a csv.

### Food trade

Datasource FAO.

# Data cleaning

## Assumptions for cleaning:

Why is EEZ NA made into 0?

Using data with F, an assumption is that having estimated data is better than no data at all.

# General cleaning

For fish trade data there are country names with strange characters that are causing problems for the “countrycodes” package, these are changed according to:

CÙte d'Ivoire -> Côte d'Ivoire

CuraÁao -> Curaçao

When joining map data with production data, all countries get included even if they don’t have any production of the specific crop. Therefor those countries value of production is set to 0.

## Categories

None if the data came with suitable, predefined categories so these had to be created manually.

#### Production and trade categories for crop and livestock

1. An excel sheet was created combining the “item group” (can be found under “definitions and standards” on individual data page) from crop production and trade data.
2. The following items where added manually since they occurred in the trade data but not in the metadata (“Oil, citronella”, “Oil, essential nes”, “Pyrethrum, dried”, “Rice - total (Rice milled equivalent)”, “Pyrethrum, extraction”).
3. Finally. each item was categorised into one of the following categories:

|  |  |  |  |
| --- | --- | --- | --- |
| Alcohol+Tobacco | Dairy products+eggs | Non-food | Roots and tubers |
| Animal oil | Fats | Nuts | Sugars |
| Animals | Fodder | Oilseeds | Vegetable oil |
| Beverages | Fruits | Processed foods | Vegetables |
| Cereals | Meat | Pulses |  |
| Coffee+cocoa+tea+spices | Miscellaneous | Rice |  |

## Creating data

To analyse countries some variables are created from the data gathered from the different sources. The variables created are mainly meant to capture economic and political drivers.

### Item fraction

This is calculated by dividing the weight (or value) of an individual item with the total weight (or value) of all items in a country. A number closer to 1 would indicate a country whose production is dominated by a few numbers of items.

### Import fraction

Calculated by dividing a country’s import by its production. This variable would indicate how much a country has to import in relation to its production. A low number would indicate that a country is not very dependent on imports.

### Export fraction

Calculated by dividing a country’s export by its production. This variable would indicate how much a country exports in relation to its production. A number close to 1 would indicate that a country exports most of its items, a number higher than 1 could indicate a trade hub.

### Import-export fraction

Calculated by dividing a country’s import by its export. This variable indicates if a country is mainly an importing or an exporting country.

## Mapping data

To map the data the main packages used was sf and tmap.

Breaks in the maps are made using “style = jenks” argument which “identifies groups of similar values in the data and maximizes the differences between categories.” (Robin Lovelace, 2019).